

СООБЩЕНИЯ
ОБЪЕДИНЕННОГО
ИНСТИТУТА
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА



14/iv-75

Ц840
А-84

11 - 8555

Д.Д.Арнаутов, Н.И.Янев

1452/2-75

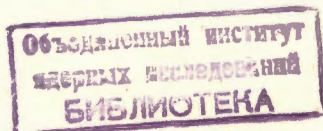
ПЕРЕВОД ИНФОРМАЦИИ И ПЕРЕКОМПОНОВКА
СТРУКТУРЫ ЗАПИСЕЙ НА МАГНИТНОЙ ЛЕНТЕ
ИНИС В ФОРМУ, УДОБНУЮ ДЛЯ ВВОДА
В ИНФОРМАЦИОННО-ПОИСКОВУЮ СИСТЕМУ ОИЯИ

1975

11 - 8555

Д.Д.Арnaudов, Н.И.Янев

ПЕРЕВОД ИНФОРМАЦИИ И ПЕРЕКОМПОНОВКА
СТРУКТУРЫ ЗАПИСЕЙ НА МАГНИТНОЙ ЛЕНТЕ
ИНИС В ФОРМУ, УДОБНУЮ ДЛЯ ВВОДА
В ИНФОРМАЦИОННО-ПОИСКОВУЮ СИСТЕМУ ОИАИ



Арнаутов Д.Д., Янев Н.И.

И - 8555

Перевод информации и перекomпоновка структуры записей на магнитной ленте ИНИС в форму, удобную для ввода в информационно-поисковую систему ОИЯИ

Описаны алгоритмы перевода информации на МЛ ИНИС в код СДС-6200. Рассмотрена методика организации и перекomпоновки информации в форму, удобную для ввода в ИПС ОИЯИ.

Сообщение Объединенного института ядерных исследований
Дубна 1975

При современном общении между различными организациями как в одной стране, так и в международном масштабе, все больше утверждаются формы обмена информацией, написанной на нестандартных носителях, таких, как перфоленты, магнитные ленты. Типичным примером такого обмена информацией в пределах одной страны является связь между региональными автоматизированными системами управления и обработки информации с подобными системами высшего иерархического уровня. Все чаще создаются и международные информационные системы в определенной области знаний, например, ИНИС /Международная система ядерной информации/. Ежемесячно ИНИС передает участвующим в ней странам и организациям копию всей обрабатываемой ею информации в записи на магнитную ленту.

Актуальным является вопрос ввода, хранения и поиска этой информации в ИПС /Информационно-поисковая система/ ОИЯИ.

Структурная схема ИПС ОИЯИ описана в работе /1/. Формат записи и организация основных информационных массивов сильно отличаются от формата записи и структуры массивов на выходной магнитной ленте ИНИС, что вызывает необходимость перекomпоновки этой структуры. Кроме того, шестибитная символьная сетка СДС /см. рис. 4/ отличается от шестибитной символьной сетки ИНИС /рис. 2/ тем, что требует предварительного перевода отдельных символов записанной на МЛ информации.

Итак, для того чтобы использовать информацию на магнитной ленте ИНИС как входную информацию в ИПС ОИЯИ, необходимо разрешить две основные проблемы:

1. Перевод символов, описывающих информацию, на МЛ ИНИС.

2. Перекомпоновку и изменение структуры записей и информационного массива на МЛ в форму, удобную для поступления на вход ИПС ОИЯИ.

Для разрешения этих проблем созданы алгоритмы, написанные на языке "ФОРТРАН".

Сначала рассмотрим весь вопрос глобально, т.е. приведем алгоритм, дающий возможность перевода информации, записанной в произвольном коде K_1 , в некоторый другой код K_2 . Изложенный ниже алгоритм осуществляет перевод произвольного символа, представленного в коде K_1 , в соответствующий символ в коде K_2 .

Пусть K_2 будет код, который используется в ЭВМ /обозначим ее через E / с размером машинного слова K битов.

Тогда $k = tk_2 + l$, где t обозначает число символов в одном машинном слове, и $0 \leq l \leq k_2 - 1$ /для ЭВМ СДС серии 6000 $k = 60, t = 10, k_2 = 6, l = 0$ /. Обозначим через ЛЦС /левый циклический сдвиг/ операцию, которая осуществляет следующую трансформацию содержимого одного машинного слова, обозначенного как p_1 / p_2 - число разрядов, на которое производится сдвиг/

для всех $i = 1, 2, \dots, k$

$i \rightarrow i - p_2$, если $i - p_2 \geq 1$, а для $i - p_2 < 1 \rightarrow i - p_2 + k$, где $i \rightarrow i^* / i^* = i - p_2$ или $i - p_2 + k$ / означает, что содержимому в i -том разряде машинного слова присваивается i^* разряд. Обычно эта операция осуществляется стандартной программой, включенной в матобеспечение ЭВМ /в СДС-6200 это программа SHIFT(p_1, p_2), где SHIFT- /сдвиг/ - имя программы, а p_1, p_2 имеют указанный выше смысл. Если $p_2 > 0$, то сдвиг будет вправо.

Для осуществления перекодировки необходима еще таблица, находящаяся в памяти E , которая задает соответствие K_1 в K_2 /пример такой таблицы приведен ниже/.

Эта таблица /в дальнейшем будем ее обозначать ТАБ/ содержит 2^{k_1} элементов / 2^{k_1} - число символов, которые могут быть представлены в коде K_1 /. j -й элемент таблицы для $j = 1, 2, \dots, 2^{k_1}$ содержит символ в коде K_2 , соответствующий символу в коде K_1 , код которого в свою очередь рассмотрен как число в десятичной системе

Таблица

1	2	3	4	1	2	3	4
элемент	символ	код	восьмеричная система				
I	пустое	55	00	42	I	II	5I
1	^	56	01	43	J	13	52
2	~	57	02	44	K	14	53
3	≡	58	03	45	L	15	54
4	≡	59	04	46	M	16	55
5	≡	60	05	47	N	17	56
6	≡	61	06	48	O	18	57
7	≡	62	07	49	P	19	58
8	^	63	08	50	Q	20	59
9	~	64	09	51	R	21	60
10	≡	65	10	52	S	22	61
11	≡	66	11	53	T	23	62
12	≡	67	12	54	U	24	63
13	≡	68	13	55	V	25	64
14	≡	69	14	56	W	26	65
15	≡	70	15	57	X	27	66
16	≡	71	16	58	Y	28	67
17	≡	72	17	59	Z	29	68
18	^	73	20	60		30	69
19	~	74	21	61		31	70
20	≡	75	22	62		32	71
21	≡	76	23	63		33	72
22	≡	77	24	64		34	73
23	≡	78	25	65		35	74
24	^	79	26	66		36	75
25	~	80	27	67		37	76
26	≡	81	28	68		38	77
27	≡	82	29	69		39	
28	≡	83	30	70		40	
29	^	84	31	71		41	
30	~	85	32	72			
31	≡	86	33	73			
32	≡	87	34	74			
33	≡	88	35	75			
34	^	89	36	76			
35	~	90	37	77			
36	≡	91	40				
37	≡	92	41				
38	≡	93	42				
39	^	94	43				
40	~	95	44				
41	≡	96	45				
	≡	97	46				
	≡	98	47				
	≡	99	50				

Замечание: В оперативной памяти содержится только 3-я колонка.

Шаг 7. $i=i+1$; если $i < t$, то следует переход к шагу 2.

Шаг 8. $A=M1$.

Шаг 9. Выход.

Пример 2.

На основании примера 1 действие этого алгоритма будет следующее:

Шаг 1. $M=7700\dots 0_{(8)}$, $M1=00\dots 0_{(8)}$, $I=1$.

Шаг 2. $C=4300\dots 0_{(8)}$.
C в код ИНИС.

Шаг 3. $M=007700\dots 0_{(8)}$,

$C=00\dots 043_{(8)} = 35_{(10)}$;
Шаг 4. $C=00\dots 003_{(8)}$;
C в коде СДС.

Шаг 5. $C=0300\dots 0_{(8)}$

Шаг 6. $M1=0300\dots 0_{(8)}$

Шаг 7. $i=2$ переход к шагу 2.
Шаги 2,3,4,5 - действия аналогичны, и т.д.

Шаг 6. $M1=031700\dots 0_{(8)}$
C \emptyset ...

Когда $i=11$, имеем в

$M1=0$ 3 1 7 15 20 25 24 05 22 23 55
C \emptyset M P U T E R S \square

Шаг 8. Содержимое $M1$ засылаем в A .

Программная реализация на ФОРТРАНе

Данные алгоритмы могут быть легко реализованы на алгоритмичном языке ФОРТРАН или PL-1 . Выбор этих языков здесь определен как из-за их широкой популярности среди потребителей ЭВМ, так и благодаря возможности осуществления логических операций.

Ниже представлены программы на языке ФОРТРАН, которые реализуют алгоритмы $A1$ и $A2$ при следующем предположении: $k=60, t=10, k_1=k_2=6$.

Через $C1, C2\dots C64$ обозначены однопозиционные символы в коде K_2 .

Программа 1.

DIMENSION ITAB (64), IT (10)

- 1.
2. DATA (ITAB (I), I=1,64) (1HC1, 1HC2...1HC64)
3. IM=7700...0B
4. ID 1 I=1,10
5. IC=IA, AND. IM
6. IC=SHIFT (IC, 6 * I)
7. IM=SHIFT (IM, 54)
8. IT (I) = ITAB (IC +1)
9. 1 C O N T I N U E

Программа 2.

- 1.
2. Как и в программе 1.
- 3.
4. IM1=0...0B
5. D \emptyset 1I=1,10
6. IC=IA. AND.IM
7. IM=SHIFT (IM, 54)
8. IC=SHIFT (IC, 6 * I)
9. IC=ITAB (IC+1)
10. IC=SHIFT (IC, 60-6 * I)
 - 1 IM = IM1 . \emptyset R . IC
12. IA=IM

Если в матобеспечении ЭВМ отсутствует программа, осуществляющая операции ЛЦС /см. SHIFT /, то указанные выше программы следует немного изменить. Для этого нужно провести операцию деления, т.е. чтобы сдвинуть содержимое слова А на р разрядов вправо /влево/, надо разделить /умножить/ А на 2^p . Например, если $A = 100100_{(2)} = 36_{(10)}$, то $A/2^2 = 36_{(10)}/4_{(10)} = 9_{(10)} = 001001_{(2)}$, т.е. содержание А после деления на 2 сдвинулось на 2 разряда вправо/. В таком случае шаг 3 из А1 надо заменить на:

$$\text{Шаг 3 } C = C/2^{k-k_1 i}, M = M/2^{k_1 i},$$

$$\text{а в алгоритме } A_2: \\ \text{Шаг 3 } M = M/2^{k_1 i}, C = C/2^{k-k_1 i}.$$

Поступающая в ИПС информация /далее используем термин "Входной массив"/ состоит из переменной длины, первые четыре символа которой содержат длину физической записи /рис. 1/. Под физической записью подразумевается информация, находящаяся между двумя последовательными промежутками на магнитной ленте.

Информация записана в коде K_1 . Для конкретности будем считать, что это код ИНИС /рис. 2/. На рис. 3 показано соответствие между дорожками магнитной ленты и 6-битным кодом ИНИС. Код K_2 будет 64-значным кодом СДС.

Для чтения записи переменной длины вообще можно использовать программу, написанную на языке КОБОЛ. Для этого достаточно указать в выражении DEPENDING ON имя /длину/ того поля записи, в котором находится число символов /в нашем примере это первые 4 символа каждой записи/. Единственная трудность здесь /и, пожалуй, непреодолимая/ состоит в следующем: транслированная программа не осуществляет перевода из одного кода в другой, и поэтому информация, которая содержится в области памяти, обозначенной в нашем примере как ДЛИНА, для СДС-6200 вовсе не будет числом /см. рис. 2,4/.

В ФОРТРАНе для осуществления ввода существует два типа оператора READ :

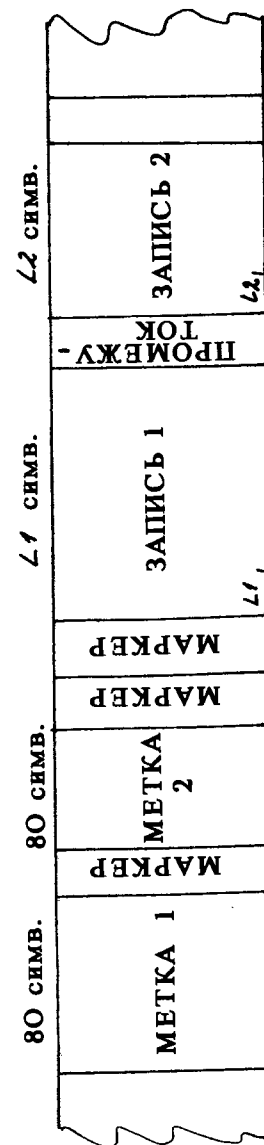
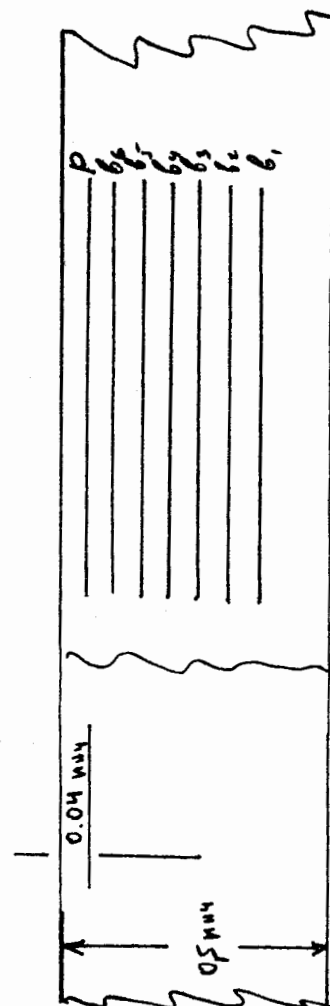


Рис. 1. Организация входного массива на магнитной ленте.

Код ИНИС - 6 битов.

	b_6	b_5	b_4	b_3	b_2	b_1
	0	0	I	I		
	0	I	O	I		
0000	L	0		P		
0001		I	A	G		
0010		2	B	R		
0011		3	C	S		
0100	\$	4	D	T		
0101	%	5	E	U		
0110		6	F	V		
0111	↑	7	G	W		
1000	(8	H	X		
1001)	9	I	Y		
1010	*	:	J	Z		
1011	+	;	K	[
1100	,	<				
1101	-	=	M]		
1110	.	>	N	^		
1111	/		O	~		

Рис. 2.



Р используется для контроля по четности /нечетности/.

Рис. 3.

- а/ READ с форматом,
- б/ READ без формата.

Первый из этих операторов нельзя использовать в данном случае по трем причинам:

1. Информация поступает записанной в коде, не совпадая с тем, который использует конкретная ЭВМ.
2. Число считанных символов зависит от списка переменных, включенных в оператор READ, а это для чтения записи переменной длины непригодно.
3. Тот бит, который на рис. 3 обозначен через Р, используется для контроля по четности или нечетности. Если в поступающей информации используем контроль по четности /нечетности/, а в конкретной ЭВМ - наоборот, тогда все символы /при нормальной работе устройств ввода/ будут объявлены ошибочными.

Последние две причины относятся также и к бесформатному READ .

Итак, для решения задачи ввода в данном примере необходим оператор ввода, который осуществляет чтение одной физической записи, без редактирования вводимой информации и с возможностью переключения по четности или нечетности.

Для СДС-6200, например, это оператор BUFFERIN(i,p), где i - номер логического устройства ввода, p=0 или 1 /контроль по четности или нечетности/.

Для совершения операции ввода необходимо еще указать начало и конец области оперативной памяти, где будет записана вводимая информация.

Итак, полная программа для осуществления ввода информации с магнитной ленты, записанной в коде ИНИС, и ее перекодирование в 64-значный код СДС будет выглядеть так /в программе используется A1 для перекодирования/:

```

1. SUBROUTINE INIS
2. COMMON IA(205), IT(2048)
3. DIMENSION ITAB(64)
4. DATA (ITAB(I), I=1,64) 1H 1HV, 1H ≠ 1H ≡ 1H$, 1H%,
5. 1H , 1H ↑ , 1H(, 1H) , 1H* ,
6. .
7. .

```

```

8. .
9. .
10. .
11. IM = 77 00000000000000000000B
12. DØ 2 j=1,205
13. 2 IA(j) = 0
14. BUFFERIN(1,1)(IA(1), IA(205))
. IF (UNIT(1).EQ.0) GØ TØ 99
.
. DØ 3 j=1,205
.
. DØ 4 j=1,10
.
. IC=IA(j).AND.IM
.
. IC=SHIFT(IC,6*1)
.
. IM=SHIFT(IM,54)
.
. 4 IT(10*(j-1)+1) = IC
.
. 3 CØNTINUE
.
. DØ 5I=1,2048
.
. 5 IT(I) = ITAB(IT(I)+1)
.
.
26. 99 RETURN

```

Остальная часть программы включается при достижении конца входного массива.

После использования BUFFERIN для устройства, и прежде чем использовать введенную информацию, состояние операции BUFFERIN должно быть проведено с использованием оператора IFUNIT. Он проверяет, действительно ли ввелись данные в ЭВМ /см. операторы 14 и 26 в приведенной выше программе/.

В рассмотренной программе принято, что максимальная длина записей массива - 2048 знаков. Так как первые четыре символа каждой записи - это длина, то не существует проблем, связанных с интерпретацией информации в записи или, точнее, с определением конца этой информации. Единственное неудобство состоит в том, что каждая

отдельная цифра числа /длина записи/ вследствие перевода расположена в отдельной ячейке памяти. Например: если $L = 1240$, то после работы подпрограммы ИНИС в ячейках ИТ(1), ИТ(2), ИТ(3), ИТ(4) оперативной памяти ЭВМ СДС будем иметь /см рис. 4/:

ИТ(1) 33333333333333333334

ИТ(2) 33.....335

ИТ(3) 3.....337

ИТ(4) 3.....33

Чтобы получить длину в 1240 символов в ячейке L /типа INTEGER /, то можно использовать операторы ENCODE|DECODE с соответствующим форматным списком /с этими операторами можно совершать перемещение информации внутри памяти вместе с редактированием информации аналогично WRITE READ с форматом/. Для указанного примера это можно сделать так:

ENCODE (10, 1, L1)(ИТ(I), I=1, 4)

DECODE (10, 2, L1)L

1 FORMAT (10A1)

2 FORMAT (I4).

Вследствие этих операций в L действительно, будет число 1240, и его можно использовать как данное типа INTEGER.

После окончания алгоритма перевода используется другой алгоритм, написанный на КОБОЛе /он здесь не приложен/, который переформирует структуру записей ИНИС /уже переведенных/ в форму, удобную для ввода в ИПС ОИЯИ. Как результат работы этого алгоритма, создается массив на МЛ СДС с двумя видами записей: первый вид имеет фиксированную длину, второй - переменную.

В этих записях соответствующим образом расположена информация, полученная с МЛ ИНИС и подлежащая вводу в ИПС ОИЯИ.

Алгоритмы реализованы на СДС-6200 и имеют следующую характеристику:

Для перевода и перекомпоновки 4000 физических

СДС - 64-значный код

ЗНАК	КОД	ЗНАК	КОД
пустое	55	Н	10
≡	74	√	66
%	63	И	11
Е	61	Ј	12
↵	65	К	13
≡	60	Л	14
^	67	М	15
↑	70	Ν	16
↓	71	О	17
>	73	Р	20
≡	75	Q	21
┘	76	Р	22
·	57	Ј	62
)	52	С	23
;	77	Т	24
+	45	И	25
\$	53	У	26
*	47	W	27
-	46	Х	30
/	50	У	31
!	56	Z	32
С	51	5	00
≡	54	0	33
≠	64	1	34
<	72	2	35
A	01	3	36
B	02	4	37
C	03	5	40
D	04	6	41
E	05	7	42
F	06	8	43
G	07	9	44

Рис. 4

записей ИНИС /около 6000 библиографических записей/ в запись этой информации на МЛ СДС, необходимо 34 мин.

Итак, для того чтобы иметь возможность достаточно эффективно формировать основные информационные массивы системы ОИЯИ, созданы алгоритмы преобразования структуры записей и массива с МЛ ИНИС на МЛ СДС-6200.

Покажем структуру получаемого массива и формат записей, которые являются входным потоком в систему информационного поиска ОИЯИ.

Для ввода информации используется МЛ СДС, на которой записан последовательный массив, состоящий из логических записей двух типов: первый имеет фиксированную длину, второй - переменную.

Каждая запись заканчивается REC-MARK, т.е. последовательностью пустых символов и закрывающейся скобкой]
10 символов.

Формат этой записи следующий:

RN	S	T	L	BLANKS	COUNT	TAGINF	RECMARK
----	---	---	---	--------	-------	--------	---------

RN - это Reference Number, который используется для идентификации данной записи;

S - STATUS CODE

T - TYPE OF RECORD

L - Bibliographical level

Blanks - эти разряды не употребляются.

17-20-COUNT/даёт число знаков записи, следующей за COUNT 1 /, который всегда равен 220.

21 - 50 - Tag 1

51 - 80 - Tag 3

81 - 110 - Tag 4

111 - 140 - Tag 5

141 - 170 - Tag 6

171 - 200 - Tag 7

201 - 230 - Tag 8

231 - 240 - Rec-Mark =]

Все значения полей указаны в работе^{/3/}. Этот тип записей содержит ровно 240 символов. Для каждого поля отведено 30 символов /пока самое большое поле - это "Tag 1", оно занимает не больше 26 символов/. Кроме того, оставлено место на случай, если на выходной МЛ ИНИС появятся и некоторые новые поля /напр. 003, 005, 007/, которые пока не встречаются в выходной информации ИНИС. Запись заканчивается группой символов, образующих поля REC-MARK. Это поле указывает на конец записи.

Во втором типе записей собирается переменная информация, т.е. информация полей с переменной длиной на МЛ ИНИС.

В конце записи также находится REC-MARK. Формат записи следующий:

RN	LEVEL	TAG	BLANK	COUNT	VARIABLELENGTH	RM
----	-------	-----	-------	-------	----------------	----

RN - Reference Number; LEVEL - Bibliographical level; TAG - это номер поля /напр., 100/, чья информация располагается в VARIABLE LENGTH.

BLANK - эти разряды не используются.

COUNT показывает число символов в следующей части записи.

RM-REC-MARK, указывающая на конец записи.

Так как в информационные массивы различные поля записей входят в различных структурах, то это определяет и предполагаемый формат входных записей. Их принадлежность к определенной библиографической записи идентифицируется RN.

Для чтения записей с МЛ СДС используется поле REC-MARK, т.е. читаются все символы до появления REC-MARK. Таким образом, можно работать последовательно с информацией каждого поля, что даёт возможность эффективно формировать информационные массивы ИПС ОИЯИ.

Показан фрагмент программы, точнее, описание входного массива на КОБОЛе.

FD FBJBL
RECORD CONTAINS 21 TO 1550 CHARACTERS
DEPENDING ON RECORD-MARK
LABEL RECORD OMITTED
DATA RECORD IS REC1 REC2

01 REC 1.
02 RN PIC 9(6).
02 S PIC 9.
02 T PIC x.

.....

01 REC 2.
02 RN PIC 9(6)
02. LEVEL PIC x(3)

.....

Литература

1. Д.Д.Арнаутов. Сообщение ОИЯИ, 10-7949, Дубна, 1974.
2. IAEA - INIS - 9, Vienna, 1972.
3. Д.Д.Арнаутов, В.А.Бирюков. Деп. публ. ОИЯИ, Б4-11-8553, Дубна, 1975.

Рукопись поступила в издательский отдел
24 января 1975 года.